Scientific and Technical Report

Multivariate Nonparametric Statistical Techniques for Simulation Model Validation

Contract No. DAAL01-96-C-0063

Data Item No. A002

Prepared for:

DEPARTMENT OF THE ARMY
U.S. Army Research Laboratory
Information Science and Technology Directorate
Building 394, Room 207
Aberdeen Proving Ground, MD 21005-5066

Attn: Ms. Ann E. M. Brodeen, AMSRL-IS-CI

Prepared by:

Edward C. Larson and B. Eugene Parker, Jr., Ph.D.

BARRON ASSOCIATES, INC.
Jordan Building
1160 Pepsi Place, Suite 300
Charlottesville, VA 22901-0807
Telephone: 804-973-1215

and

H. Vincent Poor, Ph.D.

PRINCETON UNIVERSITY
Department of Electrical Engineering
School of Engineering and Applied Science
Princeton, NJ 08544-5263
Telephone: 609-258-1816

October 21, 1997

Foreword

This document is the Final Technical Report for Contract No. DAAL01-96-C-0063, an SBIR Phase I Award funded by the U.S. Army for the nine-month period from 21 June 1996 to 20 March 1997. The research effort, which was based on the 7 July 1995 DoD SBIR Program Solicitation 95.3 Topic A95-060 entitled "Simulation Model Validation," was carried out by Barron Associates, Inc. (BAI) of Charlottesville, Virginia. B. Eugene Parker, Jr., Ph.D. served as Principal Investigator for BAI. Mr. Edward C. Larson, also of BAI, conducted most of the data analysis documented herein, and was a primary contributor to this report.

The authors express their appreciation to Malcolm S. Taylor, Ph.D. (now retired), the Government Technical Representative, Ms. Ann E.M. Brodeen, and Barry A. Bodt, Ph.D., for their helpful guidance and encouragement of BAI's efforts. The collaboration of Dr. H. Vincent Poor of Princeton University, who served as a consultant in this project, is gratefully acknowledged.

Contents

List of Figures

List of Tables

Abstract

This report documents the findings of an Army SBIR Phase I study on multivariate nonparametric tests for stochastic model validation. We herein introduce a method for generalizing rank transformations to the multivariate domain such that the rank-transformed set is uniformly distributed in multiple dimensions. This furnishes a more robust hypothesis testing technique than earlier proposed approaches and has certain computational advantages. This approach is well adapted for continuous-output models. For tests based on partitioning the model output space into bins and computing a confidence statistic based directly on bin counts, as opposed to computing statistical moments, we introduce a log-likelihood statistic that appears to be an excellent summary indicator of correspondence between a simulation model and test data. The approach is extremely versatile and well-adapted to discrete-output models.

1 Introduction

The present Phase I SBIR study is concerned with multivariate, nonparametric statistical techniques for validating stochastic simulation models. Model "validation" means ascertaining whether or not a computerized or analytic model of some phenomenon, within a certain domain of applicability, is in sufficiently accurate agreement, or *correspondence*, with reality, as represented by a finite set of test data [22]. Although there are a number of differing interpretations on the meaning and operational nature of simulation validation (See, e.g., [7, 9, 12, 17, 23, 24, 25]), the focus of this study is solely on statistical approaches based on testing for correspondence between a finite empirical body of test data and a large population of exemplars generated by a stochastic simulation model.

Computer-driven simulation of stochastic systems is a fundamental analytical technique used throughout engineering, and the natural, physical, and social sciences.[†] Thus, the development of inference-based techniques for model validation is of widespread interest. Consequently, there has been considerable attention devoted to this problem over the past three decades. (See [8] for an extensive annotated bibliography on this subject.) Prior to this effort, however, work in this field, with few exceptions (e.g., see [5]), has addressed largely only *univariate nonparametric* techniques or *multivariate parametric* techniques. The extension to *multivariate nonparametric* techniques, which is discussed herein, is nontrivial [19].

Multivariate-output, or *multiple-response*, simulation models [6] are of interest in a wide variety of applications, such as in assessing the performance of communications networks. Validating multivariate models is much more complicated than validating univariate models because of potential dependencies, or correlations, among the various output variables. An organized statistical approach to the problem of validating *stochastic univariate* simulation models was proposed and explored in [20, 21]. Since applying univariate techniques separately to each output variable fails to detect such cross-couplings, development of sound multivariate testing procedures requires special thought and consideration. As is described in detail below, we have generalized the model validation approach to problems of multivariate simulation. Although a number of *parametric* methods for approaching this problem have been proposed previously (e.g., [2]), our focus is on tests that are *nonparametric* (i.e., distribution-free), in order to obviate: (1) the undesirable necessity of assuming a population model for experimental data; or (2) the requirement of ascribing an exact, analytic statistical model to account for experimental data.

The basic framework in which model validation will be considered in this study is that of *multi-sample statistical inference*. In particular, we will consider situations in which one or more empirical realizations of the phenomenon or phenomena to be modeled are available, as well as one or more realizations produced by the simulation model. Given two such sets of data, we wish to determine whether the simulated data capture accurately the behavior of the empirical data. That is, we wish to perform a test of *homogeneity*.

In the present report, we identify and highlight the key features of the few multivariate techniques, parametric and nonparametric, that have been proposed in the literature, and demonstrate how they can be generalized or modified to yield more versatile and robust test methodologies.

[†] An application of particular interest to the Sponsor is the simulation of communication networks [10], in which the analyst wishes to determine network performance – e.g., in terms of mean throughput and delay – as a function of variable control factors – e.g., message length, message arrival rate, and transmission mode (single channel or frequency hopping).

As demonstration vehicles for these advanced methods, we present a suite of case examples that include (1) a random number generator; (2) an autoregressive moving average (ARMA) process; (3) a coin-flipping exercise; and (4) a Poisson process. These simple examples represent the essential features for a considerable range of problems, often for which computer simulation is the only practical method of analyzing system behavior.

2 Classical Techniques for Model Validation

Naylor and Finger [13] have listed several of the best-known techniques for model validation:

- *Analysis of Variance* (ANOVA) – Tests the hypothesis that the mean (or variance) of data generated by a computer simulation matches that computed from an empirical test sample. Three important assumptions underscore this technique: normality, statistical independence, and a common variance.
- χ^2 *Test* – Tests the hypothesis that a sample of simulated exemplars has the same frequency distribution as the test sample. This test is relatively sensitive to non-normality and requires selection of categories for data that are suitable and unbiased.
- *Kolmogorov-Smirnov Test* – This is a nonparametric test that involves comparing the cumulative frequency distribution of the simulated and true process data. It treats individual observations separately and, unlike the χ^2 test, does not involving partitioning the data into bins. However, it employs the normality assumptions inherent in χ^2 testing and does not generalize readily to the multivariate domain.
- *Regression Analysis* – This test involves regressing the true process data on the simulated data and testing whether the resulting regression equations have intercepts that are not significantly different from zero, and slopes that are not significantly different from unity.
- *Spectral Analysis* – This involves computing second- or higher-order spectra (polyspectra) of a time series and comparing the estimates for the simulated and test data.
- *Other Techniques* – There are a host of additional potential model validation techniques, some of which have already been investigated by others, such as *factor analysis* and *Theil's inequality coefficient.* Other examples include comparison of data compression properties, comparison of data state-space reconstructions (e.g., via Taken's embedding theorem) [26], and comparison of parametric or nonparametric (e.g., neural network) model predictions, structures, or parameter values.

In summary, the important distinguishing characteristics of various alternative test procedures are the validity of the assumptions implicit in them, their sensitivity to violations of those assumptions, and their flexibility in applying to test data other than those represented by just one empirical test database (e.g., extrapolation) [24]. Most of the above techniques are either parametric or univariate in scope and are limited in their domain of applicability and the restrictive assumptions that they employ. Our emphasis herein is on development and testing of *multivariate nonparametric* techniques, which offer wide applicability with regard to such assumptions and extensions for testing of stochastic models. *Nonparametric* procedures for simulation model validation do not rely on

assumptions regarding the statistical distributions of data. As such, they are particularly valuable in establishing the authenticity of simulations in which high confidence cannot be placed in knowledge of the true distribution of empirical data, as is generally the case with real-world data. The use of *multivariate* testing obviates the need to perform multiple univariate tests, and avoids the pitfall of increased probability of significant findings due to chance alone. The use of multivariate statistics also increases the power of a test, as univariate tests disregard the covariances among the variables and hence use less of the information available about a set of observations.

3 Moment-Based Tests for Correspondence

In this section, we elaborate on a basic theme, namely the computation of moments, common to all of the tests introduced in the literature ([5, 15, 20, 21]) to date. This technique is employed in both parametric and nonparametric tests. In this and all of the subsequent sections, our objective, stated formally, is to ascertain whether a proposed stochastic model, $\mathcal{M}$, accounts for an empirical test sample, $\mathcal{T}$, of real-world data. The fundamental strategic approach of all moment-based tests is to compare certain sample statistics of $\mathcal{T}$ (i.e., sample moments) with the values that they are expected to assume under the null hypothesis that $\mathcal{M}$ accounts for the test data in $\mathcal{T}$ or, alternatively stated, that $\mathcal{M}$ and $\mathcal{T}$ are in correspondence with one another, *viz.*, $\mathcal{M} \overset{\text{D}}{\longleftrightarrow} \mathcal{T}$, where 'D' means "as a distribution."

Since $\mathcal{T}$ is only a finite sample (size N), there will generally be some discrepancy between any sample statistic (say, for concreteness, the sample mean) computed from $\mathcal{T}$ and the expected value of that statistic determined from theoretical or simulation analysis of the surmised stochastic model, $\mathcal{M}$. To see why this is so, one could go about determining the expected value of the sample mean for $\mathcal{T}$ by generating a large number of simulation samples, all of the same size as $\mathcal{T}$, and creating a scatter plot of the resulting sample means. Even in the limit of infinitely many simulation sample sets, the distribution of sample means will have finite variance. According to the Central Limit Theorem (CLT), the distribution of sample means will be approximately Gaussian with mean μ and standard deviation $\sigma/\sqrt{N}$, where μ and σ are respectively the population mean and standard deviation of the output distribution characterizing $\mathcal{M}$. It follows that a discrepancy between the computed sample mean of $\mathcal{T}$ and the expected value, μ, is "acceptable" if it is on the order of $\sigma/\sqrt{N}$.[‡] The essence of moment-based tests, in short, is to determine whether the discrepancies between sample statistics computed from $\mathcal{T}$ and their expected values are acceptable vis-à-vis the standard error of the sample statistics. Unacceptably large discrepancies are grounds for rejecting $\mathcal{M}$.

Justification for rejecting the null hypothesis, $\mathcal{M} \overset{\text{D}}{\longleftrightarrow} \mathcal{T}$, is deemed either appropriate or inappropriate based on the scope of the validation tests that have been applied. Although the failure of a model to pass a certain battery of tests may permit its definitive rejection altogether, it is never possible, strictly speaking, to proclaim a model valid based on the results of a finite number of statistical tests, no matter how extensive. One can say only that a model has survived scrutiny or that it has not. With an *infinitely large* test sample it would be possible to make definitive pronouncements in the affirmative as well as the negative, but that is not the case in practical model validation applications.

In all of the stochastic model simulation scenarios considered in this report, it is assumed that

[‡]Of course, not all tests necessarily use this type of critical region.

simulation exemplars can be generated in arbitrarily large numbers and that sufficient randomness and independence among the generated exemplars can be achieved (e.g., by using a freshly seeded random number generator for each run). As a result, a good "picture" of the output probability distribution function (p.d.f.) describing $\mathcal{M}$ can be obtained. Empirical data, by contrast, are generally scarce.

3.1 Univariate Moment-based Tests

To elucidate the main steps of applying moment-based tests, we focus on the simple case of a stochastic model, $\mathcal{M}$, whose output is a univariate Gaussian distribution of population mean μ and standard deviation σ, *viz.*

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{1}$$

where $P(x)$ is the probability density characterizing the output of $\mathcal{M}$. We shall denote such a Gaussian distribution as $\mathcal{N}(\mu,\sigma)$, and write $\mathcal{M} \stackrel{\text{D}}{=} \mathcal{N}(\mu,\sigma)$ to express the statement that the output of $\mathcal{M}$ is a $\mathcal{N}(\mu,\sigma)$ distribution. Suppose, for instance, that we are attempting to build a random number generator that nominally returns normally distributed numbers and that we wish to find out whether it is a "good" source of random numbers.

In the special case of Gaussian models, it is convenient to apply the following linear transformation to $\mathcal{T}$, *viz.*

$$z_i \equiv (x_i - \mu)/\sigma. \tag{2}$$

This way, comparing $\mathcal{T}$ to $\mathcal{N}(\mu,\sigma)$ is equivalent to comparing $\mathcal{T}_z \equiv \{z_1,\ldots,z_N\}$ to the canonical form $\mathcal{N}(0,1)$, which has zero mean and unit variance. The sample statistics that are most frequently computed in parametric tests are the sample moments, *viz.*,

$$m_k \equiv \frac{1}{N}\sum_{i=1}^{N} z_i^k \qquad k = 1,2,3,\ldots \tag{3}$$

If N is not too small, we may appeal to the CLT to argue that the probability distributions characterizing each m_k are asymptotically Gaussian with mean M_k and standard deviation S_k given by

$$M_k = \mu_k \tag{4a}$$

$$S_k = \sigma_k/\sqrt{N} \tag{4b}$$

where

$$\mu_k \equiv \int_{-\infty}^{\infty} x^k P(x)\,dx \tag{5a}$$

$$\sigma_k \equiv \left[\int_{-\infty}^{\infty} (x^k - \mu_k)^2 P(x)\,dx\right]^{1/2} = \left[\mu_{2k} - \mu_k^2\right]^{1/2} \tag{5b}$$

are the population moments. Note that $\mu_1 = \mu$ and $\sigma_1 = \sigma$. The results for the first few moments, in the canonical case of $\mu = 0$ and $\sigma = 1$, are provided in Table 1. The general results in Eqs. 4 and

5 do not depend on the specific functional form of $P(x)$. Thus, they are not restricted to the special case of Gaussian models. Results for the case of a uniform distribution over the unit interval, *viz.*

$$P(z) = \begin{cases} 1 \text{ if } 0 < z < 1 \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

are also provided in Table 1.

Table 1: M_k and S_k Values for Gaussian (left) and Uniform (right) Distribution Models

k	M_k	S_k
1	0	$\sqrt{1/N}$
2	1	$\sqrt{2/N}$
3	0	$\sqrt{15/N}$
4	3	$\sqrt{96/N}$
5	0	$\sqrt{945/N}$
6	15	$\sqrt{10,170/N}$

k	M_k	S_k
1	1/2	$\sqrt{(1/12)/N}$
2	1/3	$\sqrt{(4/45)/N}$
3	1/4	$\sqrt{(9/112)/N}$
4	1/5	$\sqrt{(16/225)/N}$
5	1/6	$\sqrt{(25/396)/N}$
6	1/7	$\sqrt{(36/637)/N}$

The inspiration for computing higher-order moments is accredited to Reynolds ([20, 21]), whose U_k and U_k^* statistics are essentially the same as $m_k - M_k$ and $(m_k - M_k)/S_k$ respectively for the special case of one test class (see remarks below). As a numerical demonstration of a univariate moment-based test, we generated a test sample, $\mathcal{T}_1$, of $N = 100,000$ exemplars using a MATLAB routine that nominally returned a sequence of numbers representing an $\mathcal{N}(0,1)$ distribution. Computed sample moments are presented in Table 2. The fourth column gives the *confidence statistic*,

Table 2: Moments of Test Sample $\mathcal{T}_1$

k	m_k	z_k	α_k
1	1.23×10^{-5}	0.0039	0.5016
2	1.0022	0.4919	0.6886
3	−0.0086	−0.7037	0.2408
4	3.0120	0.3877	0.6509
5	−0.0544	−0.5600	0.2877
6	14.9884	−0.0364	0.4855

α_k, for the computed value of m_k vis-à-vis the probability distribution, $\mathcal{N}(M_k, S_k)$, associated with it. α_k is defined as the probability that the k'th-order sample moment of an arbitrary simulation sample, $\mathcal{S}$, of size N generated by $\mathcal{M}$ will be less than m_k. Since the probability distribution for m_k is asymptotically Gaussian with mean M_k and standard deviation S_k, it follows that α_k is equal to $\text{Erf}(z_k)$, where $z_k = (m_k - M_k)/S_k$ is the z-transform of m_k vis-à-vis $\mathcal{N}(M_k, S_k)$ and

$$\text{Erf}(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2}\, dx \tag{7}$$

is the definite integral of the normalized zero-mean, unit-variance Gaussian distribution. The value $\alpha_3 = 0.2408$, for example, indicates that were other simulation samples of size $N = 100,000$ generated by a $\mathcal{N}(0,1)$ random number generator, an m_3 value less than -0.0086 would be obtained approximately 24% of the time. Extreme values of α_k (e.g., $\alpha_k < 0.01$ or $\alpha_k > 0.99$) cast doubt on the validity of $\mathcal{M}$ as a candidate model to account for $\mathcal{T}_1$. The occurrence of two or more extreme values, as a rule, often justifies rejection of $\mathcal{M}$. The actual decision, however, of whether to reject $\mathcal{M}$ depends on the outcome of a *decision algorithm*, which typically incorporates specific threshold settings on the α_k's and examines their values *in toto*. The formulation of such decision rules for practical applications is generally a complex problem and hinges on such considerations as the relative penalties ascribed to different types of erroneous decisions that may occur.

The confidence statistics in Table 2 all have nonextreme values. There is accordingly, based on the scope of the moment-based tests applied thus far, no justification for rejecting the null hypothesis that $\mathcal{M} \stackrel{\mathrm{D}}{=} \mathcal{N}(0,1)$ is a valid model that accounts for $\mathcal{T}_1$. As an example of what happens when the univariate moment-based tests are applied to non-Gaussian test data, we generated a t-distribution with $\nu = 9$ degrees of freedom and proceed to show that it is non-Gaussian. Such a t-distribution is obtained by extracting, from a $\mathcal{N}(\mu,\sigma)$ population, a sample of size $\nu + 1 = 10$. The t-statistic, defined as

$$t \equiv \frac{m - \mu}{s/\sqrt{N}} \tag{8}$$

is computed for each sample, where m and s are the sample mean and standard deviation respectively. The resulting set of t-statistics converges asymptotically to a t-distribution with ν degrees of freedom. An expression for the exact functional form of the p.d.f. may be derived analytically, *viz.*

$$T(t) = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} . \tag{9}$$

The reason for selecting a t-distribution for demonstration was that it "looks like" a Gaussian distribution in certain superficial respects: it has zero mean, finite variance, symmetry about zero, and nonzero probability density at all real values of t. It is thus a nontrivial problem to ascertain whether a simulation sample drawn from a t-distribution fails to represent a $\mathcal{N}(0,\sigma)$ distribution, where $\sigma = \sqrt{\nu/(\nu-2)}$ is the analytically computed standard deviation of the t-distribution. To create a test sample, $\mathcal{T}_2$, we generated a t-distribution ($\nu = 9$) of $N = 10,000$ exemplars and applied the moment-based tests under the null hypothesis $\mathcal{N}(0,\sqrt{9/7}) \stackrel{\mathrm{D}}{\longleftrightarrow} \mathcal{T}_2$. The results analogous to those in Table 2 are provided in Table 3, in which $m_k = 1/N \sum_{i=1}^{N} (t_i/\sqrt{9/7})^k$.

Table 3: Moments of Test Sample $\mathcal{T}_2$

k	m_k	z_k	α_k
1	0.0092	0.9207	0.8214
2	0.9860	−0.9923	0.1605
3	−0.0246	−0.6344	0.2629
4	3.8386	8.5589	1
5	−1.1908	−3.8736	5.36×10^{-5}
6	34.5354	19.3714	1

It is evident from the results in Table 3 that the failure of the higher-order moments to agree with the M_k's under the assumption of Gaussianity (on the left in Table 1) is grounds for rejecting the null hypothesis $\mathcal{N}(0, \sqrt{9/7}) \overset{\text{D}}{\longleftrightarrow} \mathcal{T}_2$. The numbers in Table 3, *in toto*, reveal convincingly the non-Gaussian character of the t-distribution test sample, $\mathcal{T}_2$. From this simple comparison demonstration, it appears that moment-based tests provide a powerful means of model validation.

3.2 Multivariate Moment-based Tests

We show herein how the moment-based tests generalize readily to the multivariate realm. For concreteness, we focus on the case of the output distribution emerging from a linear bivariate AR(1) process, *viz.*,

$$\begin{pmatrix} y_{1,k} \\ y_{2,k} \end{pmatrix} = \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix} \begin{pmatrix} y_{1,k-1} \\ y_{2,k-1} \end{pmatrix} + \begin{pmatrix} e_{1,k} \\ e_{2,k} \end{pmatrix} \tag{10}$$

in which k is an index denoting the (discretized) time, y_1 and y_2 are a coupled pair of time-series outputs, e_1 and e_2 are a pair of independent noise channels, and $\mathbf{A}$ is a matrix of constant coefficients. For concreteness, we used the following particular set of $\mathbf{A}$ values:

$$\mathbf{A} = \begin{pmatrix} 0.2 & 0.1 \\ -0.2 & 0.6 \end{pmatrix}. \tag{11}$$

If the noise processes are both $\mathcal{N}(0,1)$-distributed, the second-order static (zero-delay) moments can be shown to compute to

$$\begin{pmatrix} \overline{y_{1,k}^2} \\ \overline{y_{1,k} y_{2,k}} \\ \overline{y_{2,k}^2} \end{pmatrix} \equiv \begin{pmatrix} \sigma_{1,1}^2 \\ \sigma_{1,2}^2 \\ \sigma_{2,2}^2 \end{pmatrix} = \begin{pmatrix} 1.0609 \\ 0.0599 \\ 1.6063 \end{pmatrix}. \tag{12}$$

Since the AR process is a linear filtering of Gaussian noise inputs, it can be shown that the output vector, $(y_{1,k}\ y_{2,k})^{\mathrm{T}}$, interpreted as a static distribution, represents a bivariate Gaussian distribution with population mean

$$\underline{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{13a}$$

and variance

$$\boldsymbol{\sigma}^2 = \begin{pmatrix} \sigma_{1,1}^2 & \sigma_{1,2}^2 \\ \sigma_{2,1}^2 & \sigma_{2,2}^2 \end{pmatrix} \tag{13b}$$

the component values for which are given in Eq. 12. Note that in the multivariate realm, the population mean of a P-dimensional probability distribution is a $P \times 1$ vector and the covariance is a $P \times P$ tensor.

Given a simulation sample, $\mathcal{S}$, of $N = 100,000$ time-series exemplars generated numerically by the AR process, we wish to test the null hypothesis $\mathcal{N}(\underline{\mu}, \boldsymbol{\sigma}) \overset{\text{D}}{\longleftrightarrow} \mathcal{S}$. Since the covariance matrix for an arbitrary zero-mean probability distribution of dimensionality, P, is always Hermitean (positive definite and symmetric), it follows that $\boldsymbol{\sigma}^2$ may be diagonalized with respect to a rotated set of orthogonal coordinate axes (the *principal axes* of the distribution) in the P-dimensional state space.

If one effects a linear transformation of the state coordinates, y_1 and y_2, from the original to the rotated frame, the resulting probability density function assumes the decoupled form

$$P(y'_1, y'_2, \ldots, y'_P) = \left[\frac{1}{\Sigma_1\sqrt{2\pi}} e^{-(y'_1)^2/2\Sigma_1^2}\right]\left[\frac{1}{\Sigma_2\sqrt{2\pi}} e^{-(y'_2)^2/2\Sigma_2^2}\right]\ldots\left[\frac{1}{\Sigma_P\sqrt{2\pi}} e^{-(y'_P)^2/2\Sigma_P^2}\right] \tag{14}$$

where $\Sigma_1, \ldots, \Sigma_P$ are the square roots of the eigenvalues of σ^2, and $y'_1, \ldots, y'_P$ are rotated coordinates, *viz.*,

$$\begin{pmatrix} y'_1 \\ \vdots \\ y'_P \end{pmatrix} = \boldsymbol{\lambda} \begin{pmatrix} y_1 \\ \vdots \\ y_P \end{pmatrix} \tag{15}$$

where $\boldsymbol{\lambda}$ is a proper rotation matrix such that $\boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}} = 1$ and $|\boldsymbol{\lambda}| = 1$. In the bivariate case, the rotation matrix assumes the form

$$\boldsymbol{\lambda} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{16}$$

in which θ is the angle of rotation between the unprimed and primed coordinate systems.

Once in the rotated frame, P separate z-distributions may be obtained by rescaling the coordinate axes to normalize the variances, *viz.*,

$$\begin{pmatrix} z'_1 \\ \vdots \\ z'_P \end{pmatrix} = \begin{pmatrix} y'_1/\Sigma_1 \\ \vdots \\ y'_P/\Sigma_P \end{pmatrix} \tag{17}$$

or, in matrix form

$$\underline{z}' = \boldsymbol{\sigma}^{-1}\underline{y}' \tag{18}$$

where $\boldsymbol{\sigma}$ is the covariance tensor that appears as

$$\boldsymbol{\sigma} = \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_P \end{pmatrix} \tag{19}$$

in the primed system.

To assess the efficacy of the multivariate moment-based tests for Gaussian models, we generated a test sample, $\mathcal{T}$, of $N = 100,000$ exemplars from the AR model in Eq. 10. The noise processes were a pair of independent $\mathcal{N}(0,1)$ processes. Hence, we expect the model output, viewed as a static distribution, to be a zero-mean bivariate Gaussian process with covariance

$$\boldsymbol{\sigma} = \begin{pmatrix} 1.0609 & 0.0599 \\ 0.0599 & 1.6063 \end{pmatrix} \tag{20}$$

as in Eqs. 12 and 13b. We wish to test the null hypothesis $\mathcal{N}(\underline{0}, \boldsymbol{\sigma}) \overset{\mathrm{D}}{\longleftrightarrow} \mathcal{T}$. To apply the moment-based tests for Gaussianity, we first transform to the rotated coordinate frame in which $\boldsymbol{\sigma}^2$ is

diagonalized. The transformation to the frame in which the variance is diagonalized is represented by the rotation matrix in Eq. 15, such that

$$\begin{pmatrix} \Sigma_1^2 & 0 \\ 0 & \Sigma_2^2 \end{pmatrix} = \lambda \sigma^2 \lambda^{-1} \tag{21}$$

where $\Sigma_1 = 1.0544$ and $\Sigma_2 = 1.6128$ are the square roots of the eigenvalues of $\boldsymbol{\sigma}^2$. In this case, the rotation angle computes to $\theta = -6.20°$. The essence of the rotation is illustrated in Fig. 1, which shows the principal axes as dashed lines. The principal axes are found by rotating the original coordinate axes in the two-dimensional state space through an angle θ. For more complicated multivariate problems with three or more variables, $P-1$ angular variables are required to specify the orientation of the principal axes. The contours of constant probability density for any bivariate Gaussian distribution are a family of concentric ellipses whose principal axes are those corresponding to the coordinate frame in which $\boldsymbol{\sigma}^2$ is diagonalized.

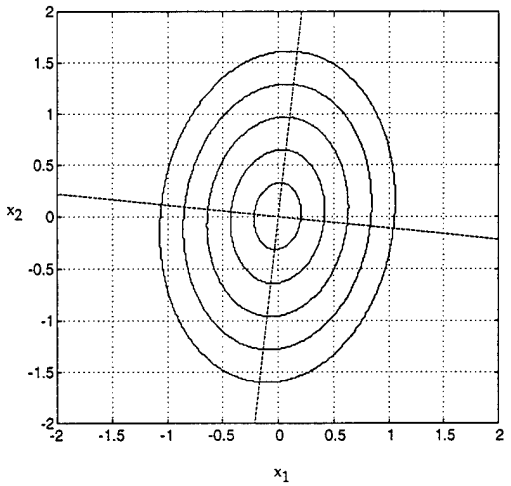

Figure 1: Elliptical Contours of Bivariate Gaussian Distribution

Having effected the transformation to the rotated frame, we apply the component-wise z-transform of Eq. 18 to obtain a transformed test sample set, which we shall denote as $\mathcal{T}_z'$. The hypothesis $\mathcal{N}(\underline{0}, \boldsymbol{I}) \overset{\text{D}}{\longleftrightarrow} \mathcal{T}_z'$ is equivalent to $\mathcal{N}(\underline{0}, \boldsymbol{\sigma}) \overset{\text{D}}{\longleftrightarrow} \mathcal{T}$, where $\boldsymbol{I}$ is the $P \times P$ identity matrix and $\mathcal{N}(\underline{0}, \boldsymbol{I})$ denotes a set of P independent zero-mean, unit-variance Gaussian distributions. The left-hand column of Table 5 provides confidence statistics for the various sample moments, m_{k_1,k_2}, of $\mathcal{T}_z'$, which are analogous to those introduced in the univariate section. For each sample moment, a z-statistic, *viz.*,

$$z_{k_1,k_2} = (m_{k_1,k_2} - M_{k_1,k_2})/S_{k_1,k_2} \tag{22}$$

and a confidence statistic

$$\alpha_{k_1,k_2} = \text{Erf}(z_{k_1,k_2}) \tag{23}$$

are computed, where

$$M_{k_1,k_2} \;=\; \mu_{k_1,k_2} = \int_{-\infty}^{\infty} x_1^{k_1} x_2^{k_2} P(x_1, x_2)\, dx_1 dx_2 \tag{24a}$$

and

$$S_{k_1,k_2} = \sigma_{k_1,k_2}/\sqrt{N} = N^{-1/2}\left[\mu_{2k_1,2k_2} - \mu_{k_1,k_2}^2\right] \tag{24b}$$

are the expected values and standard errors, respectively, of the sample moments. They are analogous to the corresponding statistics introduced in the univariate section and are related to the moments of the probability distribution function. The analytically computed results for the first few moments are provided in Table 4.

Table 4: Moments of Bivariate $\mathcal{N}(\underline{0},\mathbf{I})$ Distribution

k_1	k_2	μ_{k_1,k_2}	σ_{k_1,k_2}
1	0	0	1
2	0	1	$\sqrt{2}$
1	1	0	1
3	0	0	$\sqrt{15}$
2	1	0	$\sqrt{3}$
4	0	3	$\sqrt{96}$
3	1	0	$\sqrt{15}$
2	2	1	$\sqrt{8}$
5	0	0	$\sqrt{945}$
4	1	0	$\sqrt{105}$
3	2	0	$\sqrt{45}$
6	0	15	$\sqrt{10,170}$
5	1	0	$\sqrt{945}$
4	2	3	$\sqrt{306}$
3	3	0	15

Using Eqs. 22 and 23, we computed the moment-based test results for the hypothesis $\mathcal{N}(0,1) \overset{\text{D}}{\longleftrightarrow} \mathcal{T}_z'$. The results are displayed in the left half of Table 5.

The numbers in the left part of Table 5 all have good confidence statistics with one or two possible exceptions. Based on the scope of the moment-based tests, we cannot justifiably reject the null hypothesis $\mathcal{N}(\underline{0},I) \overset{\text{D}}{\longleftrightarrow} \mathcal{T}_z'$, or equivalently, $\mathcal{N}(\underline{0},\sigma) \overset{\text{D}}{\longleftrightarrow} \mathcal{T}$. The results thus far uphold the conjecture that the output of the AR process, when driven by Gaussian noise processes, is itself Gaussian.

As an example of a test sample that does *not* represent a Gaussian distribution, we simulated the same AR model except that the noise channels were non-Gaussian. The channels were still independent, but their outputs instead represented a uniform distribution over the interval $(-1,1)$. Note that the probability density for such noise sources is

$$P(x) = \tfrac{1}{2}\theta(x+1) - \tfrac{1}{2}\theta(x-1) \tag{25}$$

where

$$\theta(x) \equiv \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \\ \text{undefined} & \text{if } x = 0 \end{cases} \tag{26}$$

is the Heaviside step function.

Table 5: Moment-based Tests for AR Model Driven by Gaussian Noise (left) and Uniform Noise (right)

k_1	k_2	m_{k_1,k_2}	z_{k_1,k_2}	α_{k_1,k_2}
1	0	−0.0040	−1.272	0.1017
0	1	0.0037	1.161	0.8772
2	0	1.0016	−1.272	0.1017
1	1	0.0037	1.175	0.8800
0	2	0.9919	−1.811	0.0351
3	0	−0.0030	−0.248	0.4021
2	1	0.0052	0.946	0.8284
1	2	−0.0020	−0.358	0.3602
0	3	−0.0024	−0.198	0.4215
4	0	3.0423	1.363	0.9136
3	1	0.0046	0.376	0.6465
2	2	0.9874	−1.410	0.0793
1	3	0.0087	0.710	0.7611
0	4	2.9346	−2.111	0.0174

k_1	k_2	m_{k_1,k_2}	z_{k_1,k_2}	α_{k_1,k_2}
1	0	0.0013	0.396	0.6539
0	1	0.0026	0.812	0.7916
2	0	0.9991	−0.190	0.4247
1	1	−0.0001	−0.023	0.4908
0	2	0.9880	−2.673	0.0038
3	0	0.0030	0.243	0.5960
2	1	0.0022	0.401	0.6558
1	2	0.0089	1.618	0.9472
0	3	−0.0028	−0.226	0.4106
4	0	1.9355	−34.355	0
3	1	−0.0852	−6.959	0
2	2	0.9659	−3.810	6.9483×10^{-5}
1	3	0.0530	4.329	1
0	4	2.4365	−18.187	0

Equivalently, the noise channel outputs, e_1 and e_2, are described by the joint probability density function

$$P(y_1, y_2) = \begin{cases} 1/4 & \text{if } -1 < y_1 < 1 \text{ and } -1 < y_2 < 1 \\ 0 & \text{if } 0 \text{ otherwise.} \end{cases} \tag{27}$$

The variance of the uniform zero-centered distribution is

$$\overline{e^2} = \frac{1}{2}\int_{-1}^{1} x^2\, dx = \frac{1}{3}. \tag{28}$$

It is therefore fair to ask whether the output of the resulting AR process is distinguishable from that of a bivariate Gaussian process where the variances of the noise channels are both 1/3. With the assumption of such underlying noise processes, we generated another test sample of size $N = 100,000$ and transformed the AR outputs, $\underline{y}$, to the canonical form, $\underline{z}'$, using the exact same steps as above, except that the elements of $\boldsymbol{\sigma}^2$ and the eigenvalues were all scaled down by a factor of 3. Results are tabulated in the right half of Table 5.

The confidence statistics for this case are considerably poorer, especially for the fourth-order moments. Some of the z and α statistics are clearly out of the mainstream. This provides convincing evidence that the output of the AR model driven by uniform-distribution noise is non-Gaussian, i.e., the hypothesis that the test sample corresponds to $\mathcal{N}(\underline{0}, \boldsymbol{\sigma}/\sqrt{3})$. In conclusion, this shows that moment-based tests for Gaussianity extend readily to the multivariate realm and are effective in examples such as this.

An important point of note concerning the multivariate formalism introduced in this subsection is that the standard error of the first-order moments is often expressed (e.g., Eq. 9 in [15], with Eq. 8 in [15] analogous to Eqs. 12 and 13b above) as a single statistic, *viz.*,

$$V \equiv \frac{1}{N}\sum_{i=1}^{N} \frac{1}{P}(\underline{y}_i - \underline{\mu})^{\mathrm{T}} \boldsymbol{\sigma}^{-1} (\underline{y}_i - \underline{\mu}) \tag{29}$$

in which $\underline{y}$ denotes the $N \times P$ matrix of all test exemplars. This statistic, which essentially combines the $(k_1, k_2) = (0, 1)$ and $(k_1, k_2) = (1, 0)$ rows in Table 5 into a single statistic, is a direct measure of how many "standard errors of the mean" the computed sample means are from their expected values under the null hypothesis. Graphically interpreted, it indicates on which ellipse in Fig. 1 the computed joint sample mean lies.

3.3 Generalization to Multiple Test Classes

An important observation by Puri [19] concerning model validation, in general, was in noting that in many practical applications, it may be useful to develop validation tests for models that can potentially account for several different *classes* of test samples. Assigning test exemplars to different "classes" is useful and appropriate if the various test exemplars differ *a priori* in some identifiable manner (e.g., batches of data collected under different experimental conditions) that can be summarized by a set of "inputs," or control knobs $\{x_{i,1}, \ldots, x_{i,Q}\}$ in which there are Q input parameters characterizing each exemplar. Within each class (i.e., a batch of test data all having the exact same $\underline{x}_i$ input vector), it is assumed that the exemplars are i.i.d. If the input variables $\underline{x}_i$ were the same for each i (i.e., effectively ignorable), the problem reduces to one of classical homogeneity testing between the test data and a single simulation model[§] designed to account for all of the test exemplars, i.e., determining whether or not the two sets could have emerged from the same probability distribution. A number of well-known nonparametric tests can be applied to this problem (cf., [16]). The problem is somewhat more complicated, however, if there exist two or more test classes, in which case the simulation model output is different for each class, since the $\underline{x}_i$'s enter the model as a set of parameters. Each class is, in effect, a "special case" that requires generating a simulation sample tailored to it. If there are C classes, the model validation problem becomes one of validating C null hypotheses simultaneously, *viz.*,

$$\mathcal{M}_c \overset{\text{D}}{\longleftrightarrow} \mathcal{T}_c \qquad \text{for all } c = 1, \ldots, C. \tag{30}$$

Since it is intended that the larger model account for all C classes, an overall assessment of the model validity is based on how well the class-wise null hypotheses hold up on average. Puri proposes one such class average, *viz.*,

$$U_k \equiv \frac{1}{C} \sum_{c=1}^{C} z_{k,c} \tag{31}$$

in which $z_{k,c} = (m_{k,c} - M_{k,c})/S_{k,c}$, with $M_{k,c}$ and $S_{k,c}$ specific to class c. In [20], this is written as $U_1 = \sum_{i=1}^{n} t_i$, $U_3 = \sum_{i=1}^{n} t_i^2$, etc., in which n and t_i^k correspond to C and $z_{k,c}$ respectively and division by n is omitted.

4 Multivariate Hybrid Tests

Another theme common to several of the nonparametric tests that have been proposed ([5, 15, 20]) is *rank transforms*. We herein refer to these tests as "hybrid tests" because they involve computing moments of rank-transformed distributions. Alternative tests that we describe later also involve rank transforms, but do not involve computation of moments.

[§]A single model having different parameter settings, corresponding to different operating conditions $\underline{X}_i$, is, for purposes herein, considered to represent a different simulation model.

In the univariate domain, the rank-transform $\mathcal{R} = \{r_1, \ldots, r_N\}$ of $\mathcal{T}$ with respect to $\mathcal{M}$, is defined such that

$$r_i \equiv \int_{-\infty}^{y_i} P(y)\, dy \tag{32}$$

where $P(y)$ is the probability distribution of $\mathcal{M}$. The integral expression in Eq. 32 means that if a very large number of simulation exemplars is generated, r_i is equal to the percentage of such exemplars whose value is less than y_i.[¶] In the limit of infinitely many simulation exemplars, r_i asymptotically approaches the value of the integral expression in Eq. 32. If the simulation model is sufficiently tractable mathematically that $P(y)$ can be computed analytically, the rank transformation can be effected by direct evaluation of the integral. Otherwise, it is necessary to generate a very large number of simulation exemplars and count the number of exemplars less than each y_i.

The most important property of $\mathcal{R}$ is that it represents a uniform distribution over the unit interval $(0, 1)$ if and only if $\mathcal{M} \overset{\text{D}}{\longleftrightarrow} \mathcal{T}$. We will denote this distribution as $\mathcal{U}(0, 1)$. The rank transformation maps the correspondence problem from one domain to another. Instead of testing the hypothesis $\mathcal{M} \overset{\text{D}}{\longleftrightarrow} \mathcal{T}$ directly, we test the hypothesis $\mathcal{U}(0, 1) \overset{\text{D}}{\longleftrightarrow} \mathcal{R}$. The $\mathcal{U}(0, 1)$ character of $\mathcal{R}$ should emerge regardless of the functional form of $P(y)$. This transforms the problem to a canonical form that can be treated on a more unified footing.

Hybrid tests involve the application of moment-based hypothesis testing, where the null hypothesis is $\mathcal{U}(0, 1) \overset{\text{D}}{\longleftrightarrow} \mathcal{R}$. Just as sample moments were computed and evaluated vis-à-vis M_k and S_k in the preceding section, in which a Gaussian distribution was assumed, hybrid tests similarly involve computation of the sample moments of $\mathcal{R}$ and their evaluation vis-à-vis M_k and S_k, but with respect to a $\mathcal{U}(0, 1)$ distribution. The V and V^* quantities that Reynolds introduces (see [20]) are analogous to U and U^*, except that they represent moments of a uniform (rank-transformed) distribution.

5 Partial Multivariate Rank Transformations

Extending the univariate rank transform concept, as described above, to the multivariate domain, is not a simple matter. In this section, we describe and compare two methods that have been proposed for effecting rank transformations in the multivariate domain. In [5], Brodeen and Taylor propose one such partial multivariate rank transform. A related test was developed independently in [15]. In the following subsections, we clarify the distinction between the two approaches. We refer to these techniques as "partial" rank transforms because, even though they perform a (univariate) sort separately on each variate, they search for uniformity only with respect to a single variate. This contrasts with *complete* multivariate rank transforms, which we introduce in the next section.

5.1 Hybrid Test Method 1

In reference to Eq. 8 of [5], Brodeen and Taylor seek to ascertain whether a set of C test classes, $\{\mathcal{T}_c\}_{c \in 1,\ldots,C}$, are all in correspondence with one another, i.e., all characterized by a common p.d.f. and simulation model, $\mathcal{M}$. If $\mathcal{T}_c$ is expressed as a $N_c \times P$ matrix with elements $y_{i,p,c}$, where N_c

[¶] Elsewhere in the literature, including [15], division by N is omitted.

is the number of exemplars in $\mathcal{T}_c$ and P is the number of variates, Brodeen and Taylor amalgamate the $\mathcal{T}_c$'s vertically into a $N \times P$ matrix, where $N = \sum_{c=1}^{C} N_c$. For each test class, they examine the distribution of the variate-wise ranks $r_{i,p,c}$, which is the percentage of exemplars in the amalgamated set whose p'th variate is less than $y_{i,c,p}$. If the null hypothesis $\mathcal{T}_1 \overset{\mathrm{D}}{\longleftrightarrow} \mathcal{T}_2 \overset{\mathrm{D}}{\longleftrightarrow} \ldots \overset{\mathrm{D}}{\longleftrightarrow} \mathcal{T}_c \overset{\mathrm{D}}{\longleftrightarrow} \mathcal{M}$ is valid, it follows that the set $\{r_{i,p,c}\}_{i \in \mathcal{T}_c}$ should be distributed uniformly on the unit interval for each c and p. In other words, each test class should mix uniformly among all of the other classes on a variate-wise basis.

It is noted that the variate-wise uniformity of $r_{i,p,c}$ for each c is a necessary, but not sufficient, condition for the null hypothesis to be vindicated. Brodeen and Taylor proceed to test for variate-wise uniformity by computing the variance statistic

$$V_c = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{P} \left(\underline{r}_{i,c} - \frac{1}{2}\mathbf{I} \right)^{\mathrm{T}} \boldsymbol{\sigma}^{-1} \left(\underline{r}_{i,c} - \frac{1}{2}\mathbf{I} \right) \tag{33}$$

which is analogous to Eq. 29 herein except that the rank-transformed set, $\mathcal{R}$, has been used in place of the model outputs themselves. Since each $\{r_{i,p,c}\}_{i \in 1,\ldots,N_c}$ is hypothesized to be uniform on $(0,1)$, the means of the rank-transformed distributions should all be $1/2$. $\boldsymbol{\sigma}$ in Eq. 33 denotes the variance of the simulation output, $\mathcal{S}$, rank-transformed with respect to $\mathcal{M}$, where $\mathcal{S}$ is a large simulation run generated from $\mathcal{M}$ itself.

In applying this test, Brodeen and Taylor treat only the case of $C = 1$, i.e., one class at a time. In doing so, they generated a simulation sample, $\mathcal{S}_c$, for each c (corresponding to one of several different possible operating conditions for a communication network), amalgamated $\mathcal{S}_c$ and $\mathcal{T}_c$, the simulation model output vectors and measured test data vector, respectively, and computed the resulting $r_{i,c,p}$'s. They computed the variance statistic for each $\{r_{i,p,c}\}_{i \in 1,\ldots,N_c}$ set to test for variate-wise uniformity in the test class.

5.2 Hybrid Test Method 2

The essential difference between the method that we introduced in [15], and the Brodeen and Taylor approach described above, is chiefly a matter of how classes are interpreted and treated. In [5], the problem at hand involves one or more test samples that are conjectured to be representations of a single surmised simulation model. The validation method employed in [5] is then one of rank-transforming each class with respect to that simulation model and computing V_c for each class. In the approach of [15], we were concerned with a model validation scenario in which C different simulation models, $\{\mathcal{M}_c\}_{c \in 1,\ldots,C}$, were available to account for identifiable *a priori* differences in the various classes (e.g., different operating scenarios in a communication network). For each class, we generated a simulation set, $\mathcal{S}_c$, and computed $\boldsymbol{\sigma}_c$ (written as $\mathbf{Q}$ in [15]) as the variance of $\mathcal{S}_c$ rank-transformed with respect to $\mathcal{M}_c$. For each c, we computed a V_c statistic identical to that in Eq. 33, except that $\boldsymbol{\sigma}_c$ appears in our formulation in lieu of $\boldsymbol{\sigma}$. We then average the V_c's over all of the classes to obtain an overall variance statistic that indicates how well, on average, $\mathcal{M}_c$ accounts for $\mathcal{T}_c$. The formulation of [15] enables one to handle the extreme case in which all of the test exemplars are taken under a different operating condition.

To test the performance of the simulation model validation algorithm of [15] in an application example, we applied it to the bivariate AR validation scenario described in 3.2. This problem involved a surmised model with $\mathcal{U}(-1,1)$ noise channels and test samples, $\mathcal{T}_1$ and $\mathcal{T}_2$ ($N = 10,000$ exemplars each), generated using $\mathcal{U}(-1,1)$ and $\mathcal{N}(0,1/\sqrt{3})$ noise respectively. $C = 1$ in this

example. The expected value of V, under the null hypothesis, is unity. The computed values were 0.9965 and 0.9181 for $\mathcal{T}_1$ and $\mathcal{T}_2$ respectively. The corresponding confidence statistics, based on standard errors of $\sqrt{2/N}$, were 0.2475 and 5.7912. Based on these confidence statistics, the algorithm is effective in rejecting $\mathcal{M} \overset{\text{D}}{\longleftrightarrow} \mathcal{T}_2$, and, appropriately, does not reject $\mathcal{M} \overset{\text{D}}{\longleftrightarrow} \mathcal{T}_1$.

6 Complete Multivariate Rank Transformations

The partial multivariate rank transformation methods such as those described above do *not* map the simulation model p.d.f. onto a uniform distribution in P dimensions. Nor do they represent true generalizations of the rank transform methodology to the multivariate domain. Rather, they effect univariate transformations on a variate-wise basis. As a result, they may fail to detect couplings in the model outputs. As an example of how they could fail, consider the bivariate probability distribution

$$P(x, y) = 2x + 2y - 4xy \tag{34}$$

which is such that the strip integrals in both directions, $\int_0^1 P(x, y)\, dx$ and $\int_0^1 P(x, y)\, dy$, are independent of y and x respectively. Since this p.d.f. is invariant under the variate-wise rank transformation, the partial rank transform tests would fail to distinguish it from a uniform distribution on the unit square.

In this section, we propose a complete multivariate rank transform that *does* map an arbitrary P-dimensional p.d.f. onto a true multivariate uniform distribution. For a multivariate model, $\mathcal{M}$, characterized by a p.d.f. $P(y_1, \ldots, y_P)$ and a test sample $\mathcal{T} = \{\underline{y}_i\}_{i \in \{1,\ldots,N\}}$, the rank transform, $\mathcal{R}$, of $\mathcal{T}$ with respect to $\mathcal{M}$ is defined as the set $\{\underline{r}_i\}_{i \in \{1,\ldots,N\}}$, where $\underline{r}_i = (r_{i,1}, \ldots, r_{i,P})$ is the $P \times 1$ column vector such that

$$r_{i,1} = \int_{-\infty}^{y_{i,1}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P(y_1, \ldots, y_P)\, dy_1 \cdots dy_P \tag{35a}$$

$$r_{i,2} = \int_{-\infty}^{y_{i,2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P(y_2, \ldots, x_P | y_{i,1})\, dy_2 \cdots dy_P \tag{35b}$$

$$\vdots$$

$$r_{i,P} = \int_{-\infty}^{y_{i,P}} P(y_P | y_{i,1}, \ldots, y_{i,P-1})\, dy_P. \tag{35c}$$

The integral in Eq. 35a is computed by integrating the variable y_1 from $-\infty$ to $y_{i,1}$ and all of the $P - 1$ remaining y variables from $-\infty$ to ∞. An elaborate definition of the multivariate rank transform is necessary to account for correlations among the P variables in the output distribution. If the null hypothesis $\mathcal{M} \overset{\text{D}}{\longleftrightarrow} \mathcal{T}$ is valid, it follows that $\mathcal{U}(0,1)^P \overset{\text{D}}{\longleftrightarrow} \mathcal{R}$, where $\mathcal{U}(0,1)^P$ is the P-dimensional analog of the unit interval, i.e., the set of all ordered P-tuples such that the value of every component is greater than zero, but less than unity. For the bivariate case ($P = 2$), for example, $\mathcal{U}(0,1)^2$ is the uniform distribution on the unit square in the two-dimensional xy-plane, the probability density for which is

$$P(x, y) = \begin{cases} 1 & \text{if } 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{36}$$

Alternatively, the bivariate rank transform may be interpreted as the inverse of a filtering operation that takes a pair of independent $\mathcal{U}(0,1)$ random number generators and transforms the resulting

stream of numbers to emulate the output distribution of the model in question. If the surmised model is in fact valid, one will obtain a $\mathcal{U}(0,1)^P$ distribution irrespective of the ordering of the variate through which the integrals in Eqs. 35 are evaluated. An important advantage of the complete multivariate rank transform is that it maps the model output into a canonical form (*viz.*, $\mathcal{U}(0,1)^P$), for which the moments can be found from a table lookup (whereas $\boldsymbol{\sigma}$ has to be computed for each individual case in the partial rank transforms).

To demonstrate the application of the complete multivariate rank transform, we selected, as our hypothesized model, $\mathcal{M}$, the output of the same bivariate AR model in the preceding section, but with an independent pair of $\mathcal{U}(-1,1)$ distributions as the noise channels. The test samples were derived from the same AR process, but with an independent pair of $\mathcal{U}(-1,1)$ and $\mathcal{N}(0,1/\sqrt{3})$ distributions, respectively, in the noise channels. Both test samples were of size $N = 10,000$. Thus, we should expect the hypothesis $\mathcal{M} \overset{\text{D}}{\longleftrightarrow} \mathcal{T}_1$ to be vindicated, but the hypothesis $\mathcal{M} \overset{\text{D}}{\longleftrightarrow} \mathcal{T}_2$ to be refuted.

Since the noise inputs in the model $\mathcal{M}$ are non-Gaussian, the probability distribution of $\mathcal{M}$ cannot, as far as we know, be determined analytically, unlike in all of the preceding cases treated thus far. Thus, it is not possible to compute the r_i's by appealing to an analytic description of the model output. The probability distribution characterizing $\mathcal{M}$ can only be ascertained (approximately) by generating a large simulation sample, $\mathcal{S}$. In accordance with Eq. 35a, $r_{i,1}$ is defined operationally as the percentage of simulation exemplars in $\mathcal{S}$ whose first component is less than $x_{i,1}$. Since it is not possible operationally to compute the rank transform for $r_{i,2}$, as in Eq. 35b, with a simulation sample, $\mathcal{S}$, of finite size, it is necessary to make approximations. We generated a simulation sample, $\mathcal{S}$, of 500,000 exemplars using bivariate $\mathcal{U}(-1,1)$ noise and computed the rank transforms, $\mathcal{R}_1$ and $\mathcal{R}_2$, of $\mathcal{T}_1$ and $\mathcal{T}_2$ respectively vis-à-vis $\mathcal{S}$ through the following steps. The output space was partitioned into a 10×10 grid, as shown in Fig. 2. The shaded strip pertains to those test and simulation exemplars for which the computed $r_{i,1}$ value lies in the third decile, *viz.*, $0.2 \leq r_{i,1} < 0.3$. A simulation exemplar is assigned to the strip if the x_1 value of that exemplar lies in the third decile with respect to the entire simulation sample, $\mathcal{S}$. A test exemplar is assigned to the strip if between 20% and 30% of simulation exemplars assume smaller x_1 values. The finite width of this strip provides a sufficiently large subset of exemplars (on average, 1/10 of $N = 10,000$, or $1,000$) such that the distribution of x_2 values within this strip can be observed. For all points in the shaded strip, $r_{i,1}$ was "snapped" to 0.25, corresponding to the x_1-centroid of the strip. $r_{i,2}$ for each test exemplar in the strip was computed by counting the percentage of simulation exemplars in the strip that assume smaller x_2 values. In doing so, the resulting $r_{i,2}$ values are snapped to the x_2-centroid values of the corresponding deciles. The resulting set of $(r_{i,1}, r_{i,2})$ values serves as a good working approximation of the rank transform that would, in principle, be obtained from Eqs. 35 were a virtually infinite simulation sample, $\mathcal{S}$, obtainable practically. As the number of simulation exemplars increases, the grid mesh can be made finer, and thus a more accurate approximation of the exact rank transform can be obtained.

A graphical plot of the rank-transformed distributions are shown in Fig. 3, from which it is evident visually that the points in the plot for $\mathcal{T}_1$ (on the left) appear to be more uniformly dispersed than in the plot for $\mathcal{T}_2$ (on the right). In the latter, there is a preponderance of points in the center.

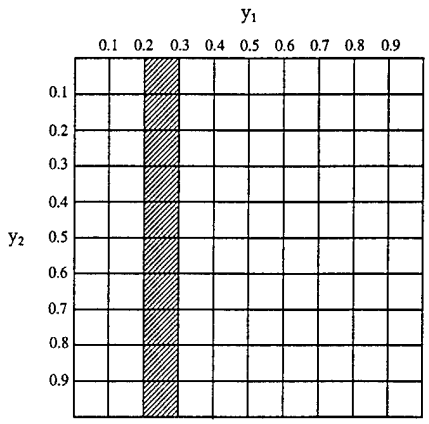

Figure 2: First Steps in Computing a Rank Transform by Cutting Strips

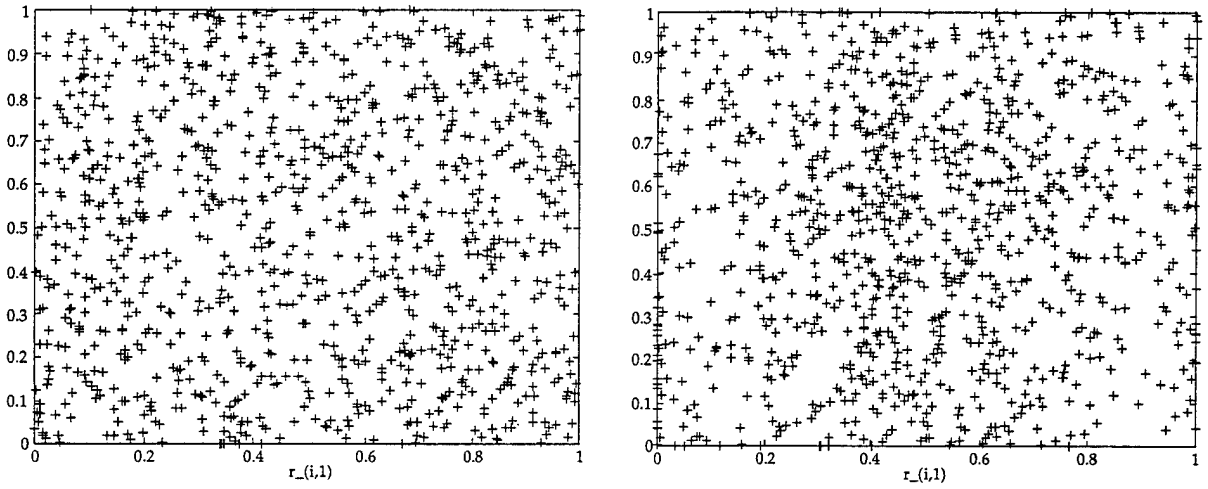

Figure 3: Rank-transforms of $\mathcal{T}_1$ (left) and $\mathcal{T}_2$ (right)

With the approximate rank transforms, $\mathcal{R}_1$ and $\mathcal{R}_2$, computed thusly, sample moments of $\mathcal{R}_1$ and $\mathcal{R}_2$ were computed. The results are displayed in Table 6. For each sample moment, m_{k_1,k_2}, the z-statistic, $z_{k_1,k_2} = (m_{k_1,k_2} - M_{k_1,k_2})/S_{k_1,k_2}$ is provided, where $M_{k_1,k_2} = \mu_{k_1,k_2}$ and $S_{k_1,k_2} = \sigma_{k_1,k_2}/\sqrt{N}$ are computed analytically. The results for the first few moments are given in Table 7. The confidence statistic, $\alpha_{k_1,k_2} = \mathrm{Erf}(z_{k_1,k_2})$, is defined as in the earlier moment-based tests.

Table 6: Hybrid Tests for AR Model Driven by Gaussian Noise (left) and Uniform Noise (right)

		z_{k_1,k_2}			
		10×10		50×50	
k_1	k_2	y_1-cut	y_2-cut	y_1-cut	y_2-cut
1	0	0.3291	1.7251	0.1732	1.8041
0	1	1.5935	0.0624	1.8755	0.1891
2	0	−0.0856	1.3016	0.0273	1.6871
1	1	1.3101	1.2219	1.3790	1.3583
0	2	1.2023	−0.2969	1.7785	0.0545
3	0	−0.2635	1.0130	−0.0152	1.5844
2	1	0.7158	0.9452	0.9085	1.3200
1	2	1.0052	0.6738	1.3824	0.9520
0	3	0.9358	−0.4277	1.6844	0.0409
4	0	−0.4214	0.7230	−0.0173	1.4985
3	1	0.3736	0.6698	0.6448	1.2213
2	2	0.5766	0.5620	0.9902	0.9950
1	3	0.7167	0.4033	1.2967	0.7439
0	4	0.6655	−0.5546	1.6086	0.0704

z_{k_1,k_2}			
10×10		50×50	
y_1-cut	y_2-cut	y_1-cut	y_2-cut
−0.8799	3.0415	−1.0905	3.0844
3.0068	−1.0115	3.0567	−1.1702
−5.2946	1.4271	−4.6738	1.8842
2.1589	1.9358	1.4823	1.9431
1.3566	−5.4260	1.9866	−4.7855
−7.8424	0.3607	−6.7373	1.0707
−1.9633	1.0890	−2.1226	1.4105
1.2895	−2.1599	0.9772	−1.5809
0.2649	−7.9786	1.2652	−6.8348
−9.1501	−0.3418	−7.5150	0.6699
−4.6770	0.2437	−4.5128	0.7671
−1.8651	−2.0477	−1.9431	−1.3851
0.4318	−4.7686	0.4337	−3.7780
−0.4503	−9.2963	0.9357	−7.5887

Table 7: Moments of Bivariate $\mathcal{N}(0,1)$ Distribution

k_1	k_2	μ_{k_1,k_2}	σ_{k_1,k_2}
1	0	1/2	$\sqrt{1/12}$
2	0	1/3	$\sqrt{4/45}$
1	1	1/4	$\sqrt{7/144}$
3	0	1/4	$\sqrt{9/112}$
2	1	1/6	$\sqrt{7/180}$
4	0	1/5	$\sqrt{16/225}$
3	1	1/8	$\sqrt{43/1344}$
2	2	1/9	$\sqrt{56/2025}$
5	0	1/6	$\sqrt{25/396}$
4	1	1/10	$\sqrt{73/2700}$
3	2	1/12	$\sqrt{109/5040}$
6	0	1/7	$\sqrt{36/637}$
5	1	1/12	$\sqrt{37/1584}$
4	2	1/15	$\sqrt{4/225}$
3	3	1/16	$\sqrt{207/12544}$

Table 6 tabulates the z-statistic values that are obtained for $\mathcal{T}_1$ and $\mathcal{T}_2$ under four different alternative approximation methods for computing the rank transform. Whereas the preceding discussion

appealed to a 10×10 grid for illustration purposes, a different mesh size, such as 50×50, is equally admissible. Results for these two alternative grid sizes are presented. Computing the rank transform by cutting strips along y_2 first is as valid as taking cutting strips along y_1 first; results for these two alternative procedures are also tabulated in Table 6 and referred to as "y_2-cut" and "y_1-cut."

The z-statistics in Table 6 exhibit consistently moderate values for $\mathcal{T}_1$ both for 10×10 and 50×50 meshes. The salient results are insensitive to whether the y_1-cut or y_2-cut method is selected. For $\mathcal{T}_2$, on the other hand, a disproportionate number of z-statistics in all four scenarios lie beyond three standard errors. These results suggest that $\mathcal{M} \overset{\text{D}}{\longleftrightarrow} \mathcal{T}_2$ be rejected based on the scope of the multivariate hybrid test, and (correctly), that $\mathcal{M} \overset{\text{D}}{\longleftrightarrow} \mathcal{T}_1$ not be rejected. It is especially remarkable that good agreement between the computed and expected moments of the rank-transformed distributions for $\mathcal{T}_1$ was obtained for both the 10×10 and 50×50 meshes. This speaks favorably to the robustness not only of the complete multivariate rank transformation method, but also of the partition method through which the rank transform was computed.

7 Multinomial Tests

In this section, we propose a multivariate nonparametric test that involves partitioning the rank-transformed set into bins and working directly with the numbers in those bins, rather than computing moments. The method is also extremely pertinent, as will be demonstrated, for scenarios in which the model output is discrete rather than continuous. As in the χ^2 test, this approach involves partitioning the rank-transformed set, $\mathcal{R}$, into bins, and evaluating the discrepancies between the observed and expected numbers of observations per bin.

The χ^2 and Kolmogorov-Smirnov (KS) tests both appeal to χ^2 distributions to compute confidence statistics. The implicit assumption, however, is that the asymptotic conditions for application of the CLT are satisfied. This, however, is not always the case. To illustrate the point, we show herein how the χ^2 test yields an erroneous confidence statistic in the simple scenario of a coin-flipping experiment.

Suppose that a fair coin is flipped 100 times and that 55 heads are obtained. We wish to reach an informed impression, based on the results of this test, as to whether the coin is fair. This means computing a confidence statistic which, in this case, is equal to the probability of obtaining fewer than 55 heads. We note that the bin counts (i.e., number of heads and number of tails) are distributed as a binomial distribution, *viz.*,

$$P(n_H) = \frac{N!}{n_H! n_T!} p_H^{n_H} p_T^{n_T} \tag{37}$$

in which $N = 100$ is the total number of flips, n_H is the number of heads obtained, $n_T = N - n_H$ is the number of tails, and $p_H = 1/2$ and $p_T = 1 - p_H = 1/2$ are the probabilities of a single flip yielding heads or tails respectively. The functional form of $P(n_H)$ is illustrated in Fig. 4.
The confidence statistic, α, is therefore equal to the $P(n_H)$ summed from $n_H = 0$ to 54. The exact result is $\alpha = 0.8159$.

The χ^2 test, by contrast, yields a confidence statistic based on the quantity

$$\chi^2 = \frac{(O_H - E_H)^2}{E_H} + \frac{(O_T - E_T)^2}{E_T} = \frac{(55-50)^2}{50} + \frac{(45-50)^2}{50} = 1.00 \tag{38}$$

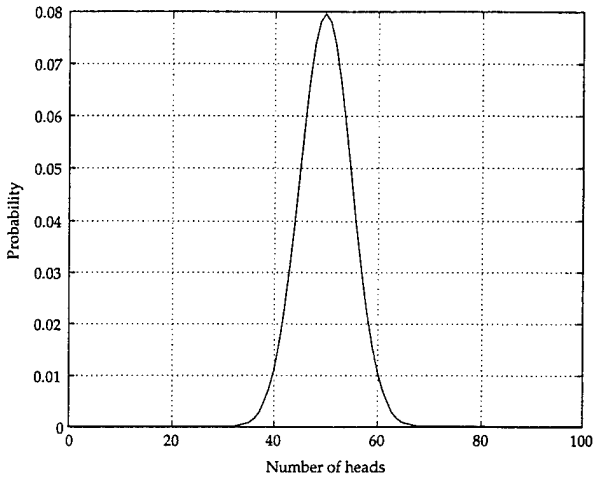

Figure 4: Probability of Obtaining n_H Heads on 100 Coin Flips

in which $O_H = 55$ and $O_T = 45$ are the observed numbers of heads and tails, and $E_H = E_T = 50$ are the expected counts. Since there are $G = 2$ bins in this case, the confidence statistic may be obtained by evaluating the definite integral $\alpha = \int_0^{\chi^2} f_\nu(v)dv$ analytically, where f_ν denotes the standard χ^2 function with $\nu = G - 1 = 1$ degrees of freedom and $\chi^2 = 1$. One obtains $\alpha = 0.6827$, which differs significantly from 0.8159. This illustrative example thus indicates that the χ^2 test, although generally very effective in determining "goodness of fit," does not necessarily yield accurate confidence statistics. The χ^2 test will yield accurate α values only if the asymptotic conditions of the CLT are satisfied.

The KS test also proves unsatisfactory in the context of this simple coin-flip scenario, but for entirely different reasons. KS applies easily only to cases in which the output of the stochastic model in question is *continuous*, as opposed to *discrete*. The rank-transformed set, $\mathcal{R}$, must assume essentially a continuum of values so that the quantity $|r'_i - i/N|$ can be computed meaningfully. Furthermore, the KS test *cannot* be extended to the multivariate domain, since there is no satisfactory way of generalizing the notion of a cumulative distribution function to probability density functions of two or more variables.

These shortcomings of the χ^2 and KS tests motivate the development of an alternative nonparametric test. The binomial distribution analysis above is the essence of such an alternative technique that we herein advocate and will refer to as the *multinomial* test. This test seeks to answer the following question: Suppose that we have a distribution consisting of $G \geq 2$ bins and n_1 observations in the first bin, n_2 in the second, ..., n_G in the G'th bin. This could represent either the output distribution of a discrete-output stochastic process or a partitioning of the rank-transformed output of a continuous-output process into G cells. Does this distribution *represent* a uniform distribution in which there are N/G observations in each bin, where $N = \sum_{g=1}^{G} n_g$ is the total number of observations?

The multinomial test models the bin counts in such scenarios as a multinomial distribution of dimension G, which states that if N observations are to be distributed among G bins, the

probability of finding n_1 observations in the first bin, n_2 in the second, etc. is equal to

$$P(n_1, \ldots, n_G) = \frac{N!}{n_1! \cdots n_G!} p_1^{n_1} \cdots p_G^{n_G} \tag{39}$$

where $N = \sum_{g=1}^{G} n_g$ is the total number of incidents and p_g is the probability of any particular incident falling into the g'th bin. For rank-transformed sets in continuous-output stochastic processes, the bin occupancy probabilities are equal, *viz.*, $p_g = 1/G$, for $g = 1, \ldots, G$. The expected value of n_g is clearly equal to $Np_g = N/G$.

The multinomial test is concerned with whether the bin counts *in toto* deviate significantly from the most probable scenario in which there are equal numbers of observations in each bin. The technique that we propose herein is to compute the log-likelihood statistic

$$\Lambda = -\ln \left[\frac{P(n_1, \ldots, n_G)}{P(N/G, \ldots, N/G)} \right] \tag{40}$$

in which $P(n_1, \ldots, n_G)$ and $P(N/G, \ldots, N/G)$ are computed as in Eq. 39. For given N and G values, the Λ statistic is characterized by a well-defined distribution whose functional form can be ascertained (approximately) through simulation. For example, the Λ distribution for $N = 100$, $G = 2$ can be examined by performing a large number of simulation runs in which a coin is flipped 100 times. In each such run, the number of heads is counted and the corresponding Λ value recorded. After a large number of such 100-flip runs, the Λ distribution emerges.

The multinomial test, like the χ^2 test, is extensible readily to the multivariate domain. The mathematical formulation is no different from in the univariate domain, and is simply a matter of working with a larger number of cells. In Section 6, we sought to test the hypotheses $\mathcal{M} \overset{\text{D}}{\longleftrightarrow} \mathcal{T}_1$ and $\mathcal{M} \overset{\text{D}}{\longleftrightarrow} \mathcal{T}_2$ by partitioning the rank-transformed sets into 10×10 or 50×50 arrays. The bin counts for the 10×10 partition for $\mathcal{T}_1$ and $\mathcal{T}_2$ are provided in Table 8. These are simply the number of exemplars that are snapped onto each cell centroid in the rank transform computation algorithm introduced in the preceding section. The upper (lower) two tables are for y_1-cut (y_2-cut) rank transforms. The second column of the first row of the upper left-hand table, for instance, indicates that for $\mathcal{T}_1$, there were 111 incidents for which $0.1 \leq \underline{r}_{i,1} < 0.2$ and $0 \leq \underline{r}_{i,2} < 0.1$. It is

Table 8: Bin Counts for $\mathcal{T}_1$ (left) and $\mathcal{T}_2$ (right)

92	111	97	95	82	83	93	99	91	120
106	105	83	96	116	89	117	104	103	93
104	97	90	117	103	100	112	82	99	109
86	94	102	114	124	90	91	102	106	106
94	89	90	107	97	112	91	102	107	95
107	99	102	92	98	98	85	124	109	90
105	95	106	95	111	87	104	117	99	108
92	92	90	106	90	114	104	108	92	99
76	96	89	107	100	107	95	98	93	109
100	123	105	91	95	98	113	96	106	98
94	100	111	77	98	101	104	96	89	80
117	104	98	98	90	100	94	89	93	136
89	85	95	97	93	98	110	89	90	96
83	102	111	113	107	89	94	104	109	97
94	109	110	124	97	104	112	93	98	92
88	91	101	94	108	95	84	112	108	98
92	116	107	90	93	85	101	115	89	110
100	107	88	99	102	124	121	96	95	96
90	106	98	111	104	105	91	97	97	122
119	99	107	97	107	84	110	104	98	96

91	85	75	97	96	92	103	91	80	84
76	66	85	63	85	90	63	70	52	63
101	76	92	113	85	98	109	89	79	102
113	110	103	120	118	126	106	110	102	129
118	98	145	140	139	151	140	146	119	127
124	92	118	134	125	161	149	148	125	150
91	120	118	117	132	131	139	134	103	133
82	75	91	83	102	125	87	108	96	94
55	53	47	73	78	75	50	54	68	62
80	75	78	100	96	114	103	84	78	79
95	77	89	103	120	133	96	86	57	68
84	68	72	107	101	91	122	85	53	86
80	85	87	107	143	118	121	91	50	66
97	56	113	122	138	130	120	83	71	88
96	85	86	119	138	126	139	102	79	109
98	99	95	137	145	150	135	124	75	95
95	58	104	110	138	149	139	92	46	112
84	72	95	105	139	141	142	103	51	99
82	54	95	103	119	122	111	82	66	78
77	87	86	124	155	139	121	83	72	79

evident visually, from casual inspection of the two sets of numbers, that the numbers in the tables

on the left are more uniform and closer to the expected value of 100 than those in the table on the right.

For both the 10×10 and 50×50 arrays, we computed the multinomial probability expressions in Eqs. 39 and 40. The Λ scores are tabulated in Table 9. Upon simulation of the Λ distribution for $(N = 10,000, G = 100)$ with 100 runs, the mean and root-variance of the distribution appear to be approximately 50 and 7 respectively, which indicate a very good confidence statistic for $\mathcal{M} \overset{D}{\longleftrightarrow} \mathcal{T}_1$, but an outlandishly large one for $\mathcal{M} \overset{D}{\longleftrightarrow} \mathcal{T}_2$. Based on the scope of the multinomial test, we can therefore reject the latter hypothesis, but the former hypothesis, i.e., $\mathcal{M} \overset{D}{\longleftrightarrow} \mathcal{T}_1$, for 10×10 bins, can clearly not be rejected based on the scope of the multinomial test. The same conclusions follow from a 50×50 partition, where a simulation of the Λ distribution for $(N = 10,000, G = 2,500)$ exhibit a mean between 1,150 and 1,200 (see Table 9).[‖]

Table 9: Λ Statistics for $\mathcal{T}_1$ (top) and $\mathcal{T}_2$ (bottom)

			Λ	α
$\mathcal{T}_1$	10×10	y_1-cut	48.38	0.42
		y_2-cut	53.53	0.70
	50×50	y_1-cut	1,231.4	~ 0.8
		y_2-cut	1,230.5	~ 0.8
$\mathcal{T}_2$	10×10	y_1-cut	357.1	1
		y_2-cut	354.2	1
	50×50	y_1-cut	1,805.2	1
		y_2-cut	1,772.6	1

7.1 Application of Multinomial Test to Bivariate Poisson Process

The log-likelihood statistic from the multinomial test offers an extremely useful and powerful technique for testing rank-transformed distributions for uniformity. The log-likelihood approach is also applicable to other types of discrete-output models because it is geared inherently toward counting discrete events in bins. As an example, we generated a bivariate Poisson process with $\lambda_A = 0.001$ and $\lambda_B = 0.003$ (i.e., λ_A and λ_B were the respective probabilities of a type A or type B event in a given time step), and ran it for 100,000 time steps.

It is well known that the probability of exactly N_A type A events and N_B type B events occurring in a duration of T time steps is $P(N_A)P(N_B)$, where

$$P(N_A) = \frac{(\lambda_A T)^{N_A}}{N_A!} \cdot e^{-\lambda_A T} \tag{41a}$$

$$P(N_B) = \frac{(\lambda_B T)^{N_B}}{N_B!} \cdot e^{-\lambda_B T} \tag{41b}$$

[‖]Phase 2 will focus on generating confidence statistics for Λ distributions in a more accurate and thorough manner.

A log-likelihood score, Λ, can be computed as

$$\Lambda = -\ln\left[\frac{P(N_A)P(N_B)}{P(\lambda_A T)P(\lambda_B T)}\right] \tag{42}$$

where $\lambda_A T$ and $\lambda_B T$ are the expected number of type A and type B events.

By generating a large number of simulation runs, the distribution of Λ can be observed. Fig. 5 shows a histogram plot for the distribution based on a simulation sample of 200 exemplars. By comparing the output of a particular test case (i.e., numbers of type A and type B events) to the distribution, a confidence statistic can be obtained and the validity of the surmised Poisson model assessed.

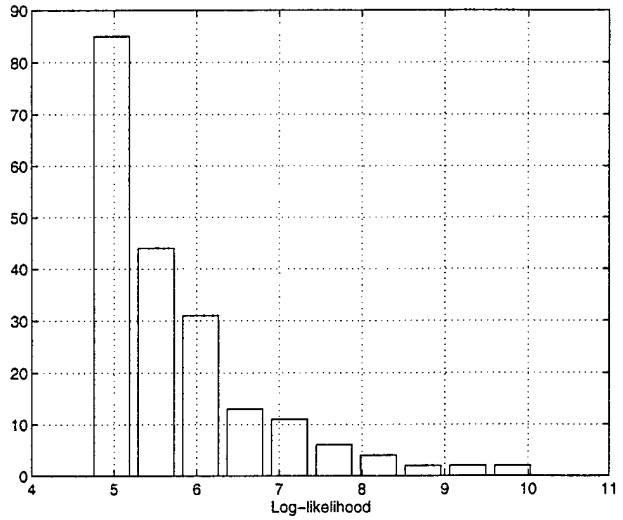

Figure 5: Simulated Distribution of Λ Values for Poisson Process

8 Conclusions

In this Phase I study, we investigated several fruitful methods of developing nonparametric tests to validate multivariate stochastic models. We began by examining, comparing, and reconciling the methods proposed earlier in [5] and [15]. Both of these methods can be classified as partial rank transformations, in that they effect essentially univariate rank transforms on a variate-wise basis. Both methods then test for uniformity in the resulting rank transformed sets by evaluating their moments. The methods differ chiefly in terms of how they interpret and treat different operating scenarios for the test and model data.

Herein, we showed how to construct a complete multivariate rank transform that maps the model output onto a uniform distribution in P dimensions. This method is a true generalization of the univariate rank transform to the multivariate domain and provides a robust test for detecting dependencies among the variates; it also converts the output distribution into a canonical form, thus obviating the need to compute the variance and higher-order moments for each and every class (i.e., operating scenario) or model.

We also showed that there are two basic ways of determining whether the model output (or a rank transformed output) conforms to the distribution predicted by the null hypothesis. One approach involves computing moments and ascertaining whether the moment values for the test sample and simulation model output are in good agreement. All of the previous attempts to construct nonparametric multivariate model validation methods espouse this basic strategy. The alternative approach is to partition the output distribution into bins and ascertain whether the observed and expected bin counts are in good agreement. We showed how the multinomial test methodology provides a more accurate correspondence assessment than the χ^2 test and how the Λ statistic serves as an excellent summary indicator of overall agreement. Both techniques appear to be extremely effective for continuous-output models, but the bin-partition approach is especially suitable for discrete-output models.

9 SBIR Phase II Effort

In Phase II, Barron Associates, Inc. will propose undertaking the following endeavors: (1) assess the relative performance and computational complexity of the various algorithms discussed herein on more elaborate synthetic test problems than was possible under the Phase I effort; (2) explore methods for developing decision algorithms (e.g., thresholds) to achieve certain specified false-positive and false-negative rates in various scenarios; and, very significantly, (3) apply these methods to a real-world stochastic, multivariate, model validation problem. Pursuant to the second goal, we plan to explore the asymptotic threshold methodology (based on the sizes of the test and simulation samples) that we suggested in [15]. The third effort will shed light on the special considerations and intricacies in applying these mathematical tools to practical problems. One such pioneering effort in this direction involving forestry data was reported in [20]. Although this study focused on univariate techniques only, its purpose was to prove the efficacy of the proposed methodology in a situation involving real-world empirical data.

As an example of a real-world stochastic, multivariate, model validation problem, consideration might be given to assessing an Army combat radio communications network, such as that discussed in [10], for which a simulation program, laboratory experimental data (conducted on a combat radio network at the Ballistic Research Laboratory), and benchmark results [5] are available. (Note that, regardless of the choice of real-world problem, the basic model validation procedure will be the same.) The purpose of the communications network experiment of [10] was to quantify the effect of three control factors, *viz.*, message length, message arrival rate, and transmission mode (single channel or frequency hopping), on network performance (defined by mean throughput and delay). These experiments were designed specifically to enable the use of statistical techniques to determine the effect of the various control factors on network performance. For this problem, a simulation of the communications network has already been coded using the commercially-available OPNET suite of tools. The communications network simulation problem is one with widespread applicability, and so simulation validation for this problem will likely be of considerable interest and relevance to the Army, as well as others.

As an alternative, BAI could readily acquire other simulation software and test data in related application areas (e.g., computer and communications networks), as well as unrelated application areas (e.g., fixed- and rotary-wing aircraft flight control systems).As a representative example of the latter, consider the validation of the NASA F-18 High Alpha Research Vehicle (HARV). At the NASA Dryden Flight Research Center (DFRC), a six-degree-of-freedom, nonlinear, batch

simulation represents one process. The other process is a hardware-in-the-loop (HIL) simulation, which includes the F-18 mission computer and flight control computer. The HIL simulation provides a more realistic test of the RFCS control law implementation for performance validation of the research flight control system (RFCS) control laws. In actuality, neither process can be considered the "true" process, as both simulations have simplifications and each can (and has) been used to discover "anomalies" with the other. Present NASA validation procedures include comparing outputs from the batch simulation and the HIL simulation for a number of test cases that specify the test flight conditions and command inputs. Plots of the two simulations are compared manually to determine if significant differences exist that may indicate anomalous or unexpected behavior. Manual validation of the HIL model against the batch simulation is very time consuming. Each anomaly discovered must be investigated and resolved by NASA engineers prior to flight testing the RFCS on the F-18 HARV aircraft. This manual review process represents a central bottleneck in the process of validating a new RFCS for the F-18 HARV, often precluding additional test cases that might otherwise prove useful.

It is noted that these simulation models may be more accurately classified as deterministic than stochastic, even though there are stochastic components due to turbulence and other noise sources (e.g., EMI). New approaches, well-founded mathematically, are needed to address such problems and would, we believe, demonstrate clear performance superiority and practicality (from the user's point of view) not only over the manual validation approach, but also over such "neural network feature-based approaches" as are currently being pursued by some (see, e.g., [1]).

As suggested by the latter application area, a potential thrust of the Phase II program might be to focus effort on the validation of models of non-stochastic systems, such as continuous-state dynamic systems, which are described by differential equations, and *deterministic* discrete-event systems, which include temporal logic, min-max algebraic models, finite state machines, and Petri nets.

Still another alternative for the Phase II effort might be to address simulation model validation problems across the most commercially-relevant classes of models, with the goal of creating a "toolbox" of procedures for users covering many types of simulation models.

References

[1] AIM System Validator (AIM SV) for the F-18 High Alpha Research Vehicle – Concept Design & Feasibility Demonstration, AbTech Corporation, NASA Phase II SBIR Contract NAS4-50033, Mar. 20, 1995 – Feb. 28, 1997.

[2] O. Balci and R.O. Sargent, "A methodology for cost-risk analysis in statistical model validation of simulation models," *Comm. ACM,* Vol. 24, No. 11, 1981.

[3] Y. Barlas, "An autocorrelation function test for output validation," *Simulation*, July 1990, pp. 7 - 16.

[4] P.J. Bickel, "Some asymptotically nonparametric competitors of Hotelling's T^2," *Ann. Math. Stat.,* Vol. 36, pp. 160 - 173, 1965.

[5] A.E.M. Brodeen and M.S. Taylor, *A Multivariate Multisample Rank Test for Stochastic Simulation Validation*, Final Technical Report to U.S. Army Research Laboratory, ARL-TR-592, October 1994.

[6] D.S. Burdick and T.H. Naylor, "Design of computer simulation experiments for industrial systems," *Comm. ACM*, Vol. 9, No. 5, pp. 323 - 329, 1966.

[7] P.N. Finley and J.M. Wilson, "The paucity of model validation in operational research projects," *J. Opl. Res. Soc.,* Vol. 38, No. 4, pp. 303 - 308, 1987.

[8] M.A. Hamilton, "Model validation: An annotated bibliography," *Commun. Statist. – Theory Meth.,* Vol. 20, No. 7, pp. 2207 - 2266, 1991.

[9] G.F. Hermann, "Validation problems in games and simulations with special reference to models of international politics," *Behavioral Science,* Vol. 12, pp. 216 - 231, 1967.

[10] V.A.T. Kaste, A.E.M. Brodeen, B.D. Broome, *An Experiment to Examine Protocol Performance Over Combat Net Radios.* U.S. Army Ballistic Research Laboratory Report No. BRL-MR-3978, June 1992.

[11] L. Ljung, *System Identification – Theory for the User.* (Prentice-Hall: Englewood Cliffs, NJ, 1987, pp. 424 - 430)

[12] G.A. Mihram, "Some practical aspects of the verification and validation of simulation models," *Operational Res. Quart.*, Vol. 23, No. 1, pp. 17 - 29, 19.

[13] T.H. Naylor and J.M. Finger, "Verification of computer simulation models," *Management Science,* Vol. 14, No. 2, pp. B92 - B101, 1967.

[14] T.H. Naylor, *The Design of Computer Simulation Experiments.* (Duke University Press: Durham, 1969)

[15] B.E. Parker, Jr. and H.V. Poor, *Statistical Techniques for Simulation Model Validation*, Interim Report for Army Contract DAAL01-96-C-0063, (Barron Associates, Inc., Oct. 1996)

[16] H.V. Poor, *An Introduction to Signal Detection and Estimation – Second Edition.* (Springer-Verlag: New York, 1994)

[17] K.R. Popper, *The Logic of Scientific Discovery.* (Basic Book: New York, 1959)

[18] M.L. Puri, "On the Combination of Two-sample Tests of a General Class," *Rev. Int. Stat. Inst.*, Vol. 33, pp. 229 - 241, 1965.

[19] M.L. Puri and P.K. Sen, *Nonparametric Methods in Multivariate Analysis.* (Wiley: New York, 1971)

[20] M.R. Reynolds, Jr., H.E. Burkhart, and R.F. Daniels, "Procedures for statistical validation of stochastic simulation models," *Forest Sci.*, Vol. 27, No. 2, pp. 349 - 364, 1981.

[21] M.R. Reynolds, Jr., and M.L. Deaton, "Comparisons of some tests for validation of stochastic simulation models," *Commun. Statist. – Simula. Computa.*, Vol. 11, No. 6, pp. 769 -799, 1982.

[22] Society for Computer Simulation – Technical Committee on Model Credibility, "Terminology for model credibility," *Simulation*, Vol. 32, No. 3, pp. 103 - 104, 1979.

[23] G. Sheng, M.S. Elzas, T.I. Ören and B.T. Cronhjort, "Model validation" A systemic and systematic approach," *Reliability Engineering and System Safety,* Vol. 42, pp. 247 - 259, 1993.

[24] R. Van Horn, "Validation," in *The Design of Computer Simulation Experiments*, pp. 232 - 251. T.H. Naylor, Ed. (Duke University Press: Durham, 1969)

[25] D. Watts, "Time Series Analysis," in *The Design of Computer Simulation Experiments*, pp. 165 - 203. T.H. Naylor, Ed. (Duke University Press: Durham, 1969)

[26] A.S. Weigend and N.A. Gershenfeld, *Time Series Prediction: Forecasting the Future and Understanding the Past.* (Addison-Wesley: Reading, MA, 1994)